
HIVE: Enabling Common Language and Interdisciplinarity

EPA-NIEHS Advancing Environmental
Health Data Sharing and Analysis:
Finding a Common Language
June 25, 2013

Jane Greenberg, Professor SILS
Director, SILS Metadata Research
Center



Overview

- Languages of aboutness
- Ontology
- Vocabulary challenge(s) *re ...* scientific data
- HIVE—Helping Interdisciplinary Vocabulary Engineering
- Conclusions, Q & A

Languages for aboutness

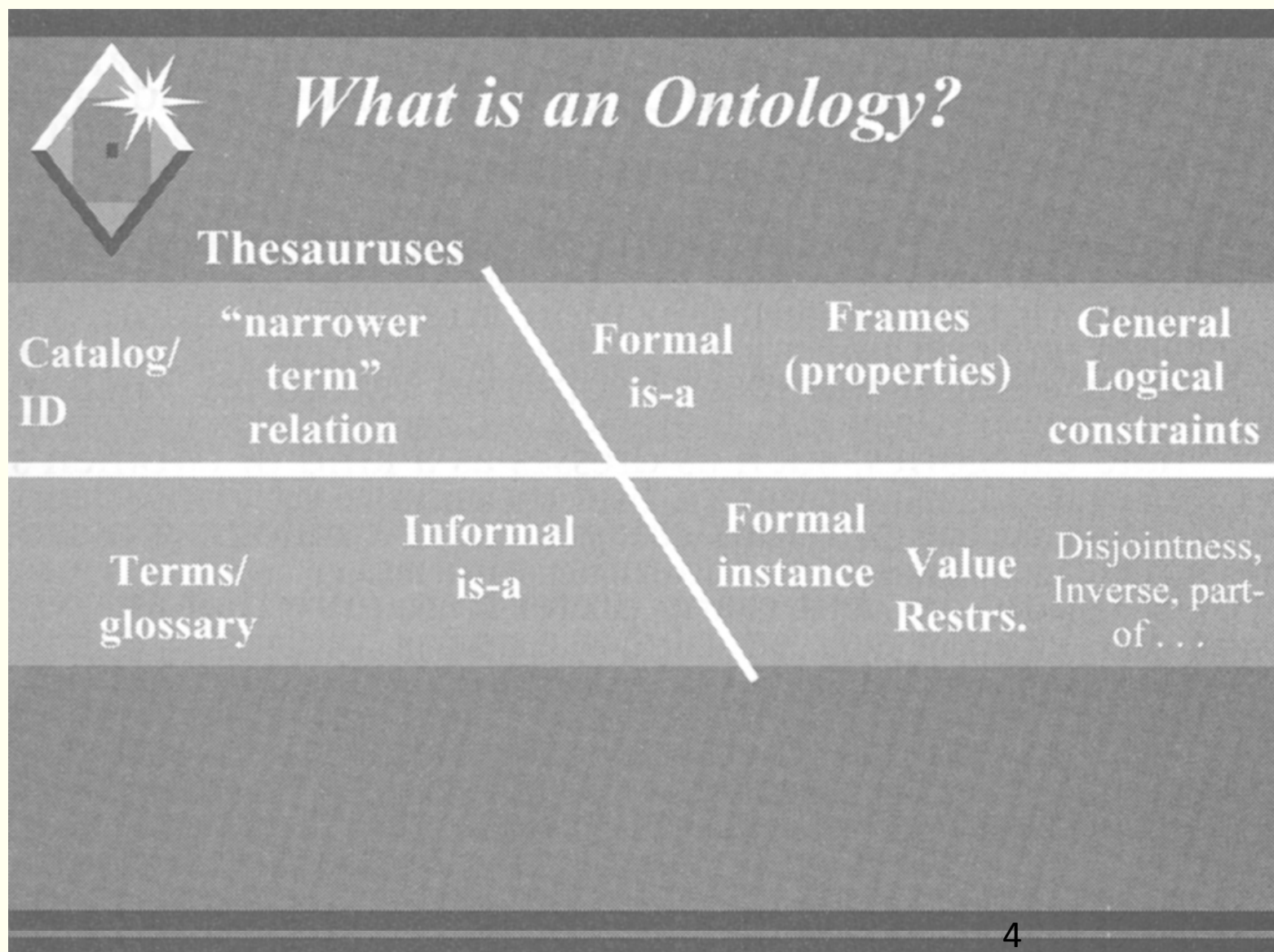
A Language

- A systematic arrangement of concepts
- What makes a language systematic?
- What makes an indexing language systematic?

Advantages & disadvantages

- Discovery
- Communication
- Interoperability
- Browsing, serendipity
- Context, grouping
- Overview of the scope of a service
- Partitioning / Segmenting (facets)
- Multilingual access
- Known by users
- Machine processing
- Costly
- Stagnant/difficulty in adding new concepts.

(McGuinness, D. L. (2003). Ontologies Come of Age. In Fensel, et al, *Spinning the Semantic Web*. (Cambridge, MIT Press), pp. 175. [see also, p. 181 + 189])



Vocabulary challenge(s) and scientific data management

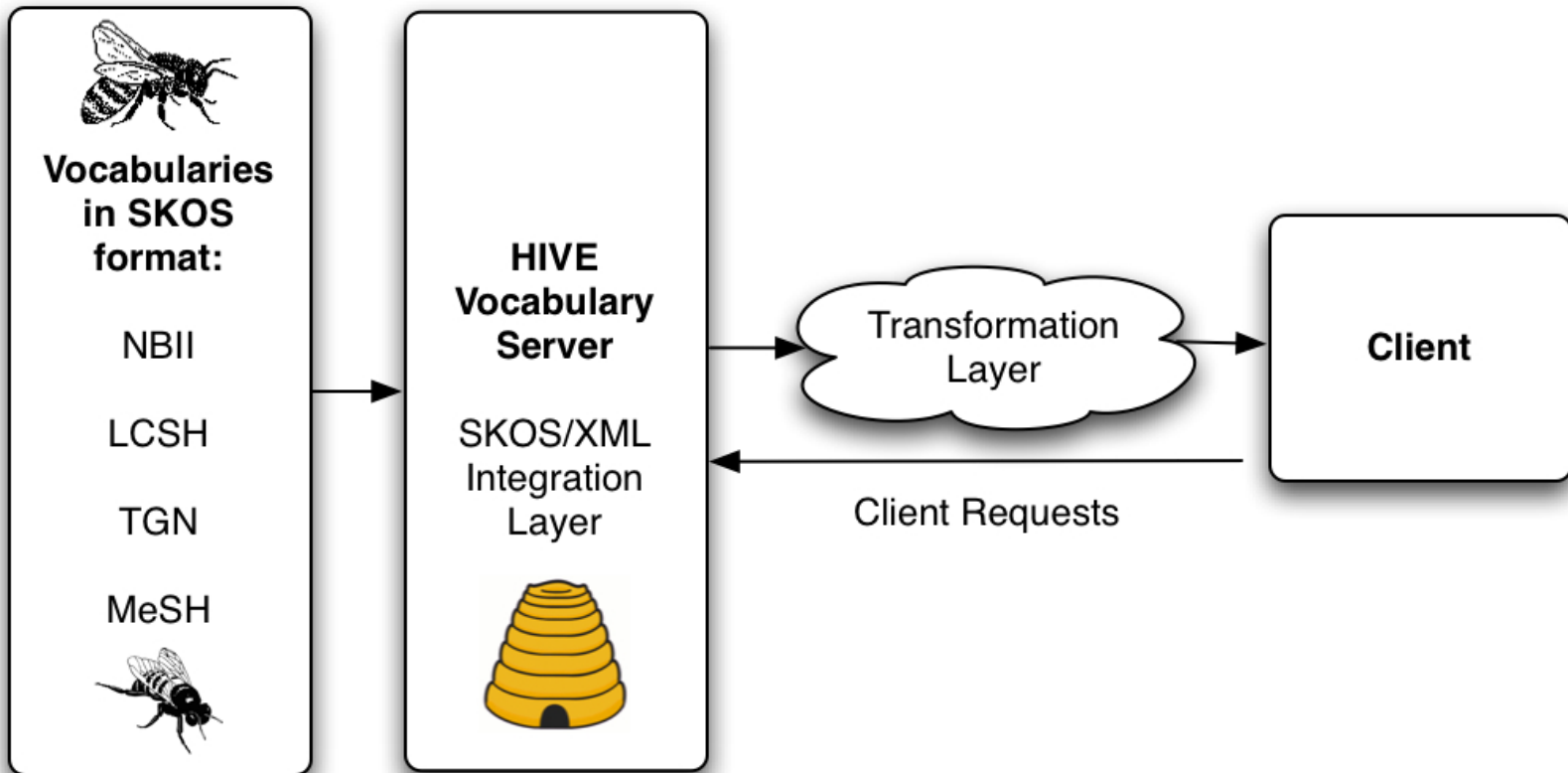
- Research Challenge

Apply standard vocabulary terms to data in collections to improve organization and discovery

- Applications needed to...

- Help researchers select appropriate terms for describing data sets
- Integrate terminology selection with data ingestion tools
- Apply standard vocabularies and not reinvent the wheel

HIVE model



- <AMG> approach for integrating discipline CVs
- Model addressing C V cost, interoperability, and usability constraints (interdisciplinary environment)

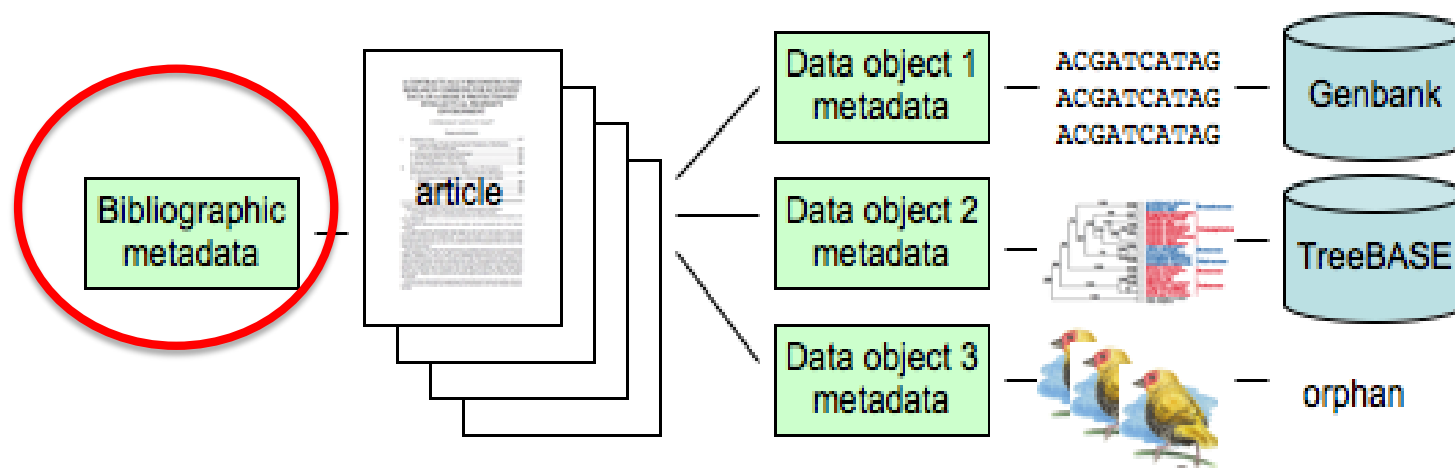


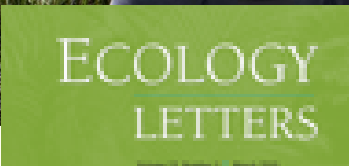
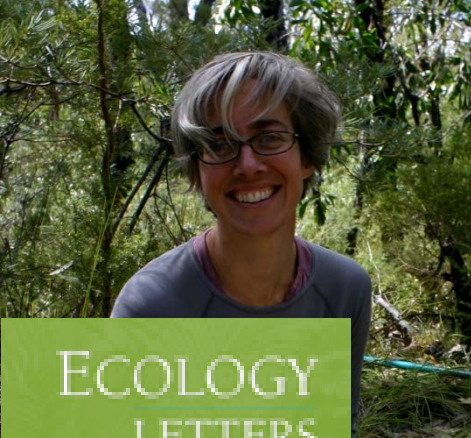
Dryad...nonprofit organization and an international repository of data underlying scientific and medical publications

Results from study with 600 keywords

- 431 **topical** terms, exact matches: NBI Thesaurus, 25%; MeSH, 18%
- 531 terms (**topical terms, research method and taxon**): LCSH, 22% found exact matches, 25% partial

Conclusion: Need multiple vocabularies





~~~~Amy

- Meet Amy Zanne. She is a botanist.
- Like every good scientist, she publishes, and she deposits data in Dryad.

| Family         | Binomial             | A (mm <sup>2</sup> ) | F (mm <sup>2</sup> /mm <sup>2</sup> ) | N (mm <sup>-2</sup> ) | S (mm <sup>4</sup> ) |
|----------------|----------------------|----------------------|---------------------------------------|-----------------------|----------------------|
| Caprifoliaceae | Abelia biflora       | 0.002375829          | 0.924197654                           | 389.0                 | 6.10753E-06          |
| Caprifoliaceae | Abelia dielsii       | 0.00115375           | 0.357418211                           | 331.0                 | 3.48565E-06          |
| Caprifoliaceae | Abelia integrifolia  | 0.001134115          | 0.240432369                           | 212.0                 | 5.3496E-06           |
| Caprifoliaceae | Abelia mosanensis    | 0.000855299          | 0.632065665                           | 739.0                 | 1.15737E-06          |
| Caprifoliaceae | Abelia serrata       | 0.000706858          | 0.206402637                           | 292.0                 | 2.42075E-06          |
| Caprifoliaceae | Abelia spathulata    | 0.000804248          | 0.230819095                           | 287.0                 | 2.80226E-06          |
| Malvaceae      | Abutilon fruticosum  | 0.001452201          | 0.137959114                           | 95.0                  | 1.52863E-05          |
| Malvaceae      | Abutilon pannosum    | 0.003117245          | 0.124689812                           | 40.0                  | 7.79311E-05          |
| Fabaceae       | Acacia albida        | 0.012271846          | 0.049087385                           | 4.0                   | 0.003067962          |
| Fabaceae       | Acacia ataxacantha   | 0.013069811          | 0.169907541                           | 13.0                  | 0.00100537           |
| Fabaceae       | Acacia borleae       |                      |                                       | 15.0                  | 0.000271434          |
| Fabaceae       | Acacia burkei        |                      |                                       | 6.0                   | 0.001498671          |
| Fabaceae       | Acacia caffra        |                      |                                       | 21.0                  | 0.000486049          |
| Fabaceae       | Acacia cyanophylla   |                      |                                       | 22.0                  | 0.000416404          |
| Fabaceae       | Acacia davyi         | 0.008332289          | 0.099987469                           | 12.0                  | 0.000694357          |
| Fabaceae       | Acacia erioloba      | 0.015174678          | 0.091048067                           | 6.0                   | 0.002529113          |
| Fabaceae       | Acacia erubescens    | 0.008824734          | 0.07059787                            | 8.0                   | 0.001103092          |
| Fabaceae       | Acacia exuvialis     | 0.001134115          | 0.018145839                           | 16.0                  | 7.08822E-05          |
| Fabaceae       | Acacia galpinii      | 0.012076282          | 0.096610257                           | 8.0                   | 0.001509535          |
| Fabaceae       | Acacia gerrardii     | 0.011574413          | 0.098023581                           | 7.5                   | 0.001543255          |
| Fabaceae       | Acacia grandicornuta | 0.006503882          | 0.045527175                           | 7.0                   | 0.000929126          |
| Fabaceae       | Acacia haematoxylon  | 0.005026548          | 0.095504417                           | 19.0                  | 0.000264555          |

Amy's data



[Home](#)
[Concept Browser](#)
[Indexing](#)

Opened vocabularies: [XNBII](#) [XLCSH](#) [XAGROVOC](#) [+Add](#)


[NBII](#)
[LCSH](#)
[AGROVOC](#)

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#)  
[N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)  
[\[0-9\]](#)

- [+ Abundance \(organisms\)](#)
- [+ Accidents](#)
- [+ Accumulation](#)
- [+ Action potential](#)
- [+ Activity](#)
- [+ Adherence](#)
- [+ Administration \(drugs\)](#)
- [+ Aeration](#)
- [+ Age \(biology\)](#)
- [+ Age \(geology\)](#)
- [+ Agents](#)
- [+ Agricultural products](#)
- [+ Agriculture](#)
- [+ Air masses](#)
- [+ Airports](#)
- [+ Algorithms](#)
- [+ Allergenicity](#)
- [+ Analysis](#)
- [+ Anesthesia](#)
- [+ Animal products](#)
- [+ Antigen-antibody complexes](#)

Your search for **wood** returns following concepts:

[AGROVOC Improved wood](#)  
[LCSH Wood--Identification](#)  
[LCSH Wood, Compressed](#)  
[LCSH Compression wood](#)  
[LCSH Wood distillation](#)  
[LCSH Wood--Deterioration](#)  
[LCSH Fireproofing of wood](#)  
[LCSH Simulated wood](#)  
[LCSH Wood--Color](#)  
[LCSH Wood--Microbiology](#)  
[LCSH Wood flour](#)  
[LCSH Wood--Research](#)  
[LCSH Wood--Utilization](#)  
[NBII Wood pulp](#)  
[AGROVOC Wood residues](#)  
[AGROVOC Wood properties](#)

Filter the result

- ☒ AGROVOC
- ☒ LCSH
- ☒ NBII

**NBII->Wood pulp**

[View in SKOS](#)

|                          |                                                                                                                         |
|--------------------------|-------------------------------------------------------------------------------------------------------------------------|
| <b>Preferred Label</b>   | Wood pulp                                                                                                               |
| <b>URI</b>               | <a href="http://thesaurus.nbii.gov/nbii#Wood-pulp">http://thesaurus.nbii.gov/nbii#Wood-pulp</a>                         |
| <b>Alternative Label</b> | Pulp (wood);                                                                                                            |
| <b>Broader Concepts</b>  | <a href="#">Wood</a>                                                                                                    |
| <b>Narrower Concepts</b> | This concept does not have narrower terms.                                                                              |
| <b>Related Concepts</b>  | <a href="#">Paper</a><br><a href="#">Pulp mills</a><br><a href="#">Sawdust</a><br><a href="#">Paper industry wastes</a> |

[Home](#)[Concept Browser](#)[Indexing](#)

HIVE vocabulary server provides functionality to identify concepts from given document or text. You need only two easy steps to get the concepts that are relevant to your document:

- Step 1: Select the vocabulary source
- Step 2: Upload your document **OR** Enter the URL of your document

### HIVE Automatic Concepts Extractor

Step 1: Select vocabulary source

☒ AGROVOC☒ LCSH☒ NBII

Step 2: Upload a document

**OR** Enter the URL

Powered by



# REVIEW AND SYNTHESIS

## Towards a worldwide wood economics spectrum

Jerome Chave,<sup>1\*</sup> David Coomes,<sup>2</sup>  
Steven Jansen,<sup>3</sup> Simon L. Lewis,<sup>4</sup>  
Nathan G. Swenson<sup>5</sup> and Amy E.  
Zanne<sup>6,7</sup>

<sup>1</sup>Laboratoire Evolution et  
Diversité Biologique, UMR 5174,  
CNRS/Université Paul Sabatier  
Bâtiment 4R3 F-31062 Toulouse,  
France

### Abstract

Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant optimizes among these competing functions can have major ecological implications, which have been under-appreciated by ecologists compared to the focus they have given to leaf function. To draw together our current understanding of wood function, we identify and collate data on the major wood functional traits, including the largest wood density database to date (8412 taxa), mechanical strength measures and anatomical

### Extracted Concepts Cloud

AGROVOC  
LCSH  
NBII

Reaction wood    Wood--Figure    Wood--Discoloration    Calavicci, AI (Fictitious character)    Lāt,  
al- (Arabian deity)    Murphy, AI (Fictitious character)    Density    Soils--Density    Population  
density    Recessive traits    Traits (genetics)    Dominant traits    Associated species    Species  
diversity    Numbers of species    Plant anatomy    Plant litter    Plant condition    Leaf  
spots    Leaf prints    Leaf blowers    Brushes, Carbon    Electrodes, Carbon    Carbon  
taxes    Growth    Fetus--Growth    Growth (Plants)    Infiltration water    Water--  
Color    Drinking water

# About HIVE...

| Goal                                                                                                                                                                                                                                                                                                                                                        | Plan                      | Vocabulary Partners                                                                                                                                                                                                                                                                                                                                                               | Workshop Hosts                                                                                                                                                                                                                                  |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>• Provide efficient, affordable, interoperable, and user friendly access to multiple vocabularies during metadata creation activities</li> <li>• Present a <b>model</b> and an <b>approach</b> that can be replicated <ul style="list-style-type: none"> <li>▪—&gt; not necessarily a service</li> </ul> </li> </ul> | Build<br>Plan<br>Evaluate | <ul style="list-style-type: none"> <li>• Library of Congress: LCSH</li> <li>• Getty Research Institute (GRI): TGN (Thesaurus of Geo. Names )</li> <li>• United States Geological Survey (USGS): NBII Thesaurus, Integrated Taxonomic Information System (ITIS)</li> <li>• National Library of Medicine and the</li> <li>• National Agricultural Library</li> <li>• FAO</li> </ul> | <ul style="list-style-type: none"> <li>• Columbia Univ.</li> <li>• Univ. of California, San Diego</li> <li>• George Washington University</li> <li>• Univ. of North Texas</li> <li>• Universidad Carlos III de Madrid, Madrid, Spain</li> </ul> |



# HIVE Team

José R. P. Agüera

Lina Huang

Lee Richardson

Bob Losee

Hollie White

Madhura Marathe

Jane Greenberg

Craig Willis

Ryan Scherle





## Helping with Interdisciplinary Vocabulary Engineering

[Home](#)[Concept Browser](#)[Indexing](#)

### Welcome to HIVE!

Helping Interdisciplinary Vocabulary Engineering(HIVE) is an IMLS funded project involving the Metadata Research Center (MRC) at the School of Information and Library Science, University of North Carolina at Chapel Hill, and the National Evolutionary Synthesis Center (NESCent) in Durham, North Carolina. Below you will find our experimental functioning HIVE system. You are welcome to try our SKOS-based system by browsing concepts from interdisciplinary vocabularies or experience a new approach to metadata generation by using the indexing feature.

#### Search a Concept

Browse and search concepts in selected vocabularies.

#### Index a Document

Automatically extract document concepts for subject metadata creation.

*This HIVE system is for demo purposes and may change in response to your feedback. [Contact us](#)*

#### Vocabulary Statistics

| Vocabulary              | Concepts | Relationships | Last Updated |
|-------------------------|----------|---------------|--------------|
| <a href="#">AGROVOC</a> | 28174    | 83086         | Jun 12, 2011 |
| <a href="#">ITIS</a>    | 391775   | 783438        | Aug 21, 2011 |
| <a href="#">LCSH</a>    | 406631   | 541623        | Aug 3, 2011  |
| <a href="#">MeSH</a>    | 25610    | 76649         | Aug 21, 2011 |
| <a href="#">NBII</a>    | 8680     | 46432         | Jun 12, 2011 |
| <a href="#">TGN</a>     | 895197   | 1799154       | Jun 13, 2011 |




Metadata Research Center <MRC>





# HIVE in LTER, Dryad,...





## EML Tagger

To view a list of suggested keywords, browse to an EML File on your system and click Tag Document

**Vocabulary Options:**

☒ LTER Controlled Vocabulary

☒ CSA/NBII Biocomplexity Thesaurus

EML File:  No file chosen

**Suggested LTER Keywords**

- marine
- oceans
- habitats
- fishes
- organisms
- water
- biology
- corals
- species
- coral reefs

**Suggested NBII Keywords**

- Seas
- Natural habitat

**Submit Data Now!**

[See how to submit](#)

**Title:** Data from: Climatic stability and genetic divergence in *Anolis krugi* from Jardinerio de la Montaña

**Abstract:** Two factors that can lead to geographic structuring of populations are geographic barriers and physiological constraints. Populations that occur in different physiographic regions may be restricted to those areas by physical barriers, which may facilitate the formation of phylogeographic clades. Long-term climatic stability of populations may also facilitate diversification, because new clades are likely to evolve in areas that experience lesser climatic shifts. We conducted a phylogeographic study of the Puerto Rican lizard *Anolis krugi* to assess whether populations of this species are structured across the species' range, and if they do, whether these breaks coincide with the boundaries of physiographic regions of Puerto Rico. We also assessed whether interpopulation genetic distances are correlated with relative climatic stability in the island. *Anolis krugi* exhibits genetic structuring, but these patterns do not correspond to the physiographic regions of Puerto Rico. We used climatic reconstructions of two eras, the Quaternary period, the present conditions and those during the last glacial maximum, to quantify climatic stability between sampling locations. We documented positive correlations between genetic distance and climatic stability, although these associations were not significant when corrected for autocorrelation. The analysis was employed to assess the relationship between climatic stability and the genetic architecture of *A. krugi*. We investigated the impact of factors such as the spatial distribution of food sources, parasites, predators, and competitors on the genetic landscape of a species.

Use this interface to add, remove, or enhance the subject and scientific name metadata for this record. Use the interface to map free-text keywords to controlled terms. The "Suggested Terms" panel displays a list of terms automatically suggested based on the resource title, abstract, and keywords.

**Keywords**

- ☒ Climate Change
- ☒ Phylogeography
- ☒ Population Genetics - Empirical
- ☒ Reptiles

**Scientific Names**

- ☒ *Anolis krugi*

**Lookup term**

**Suggested Terms ?**

- ☐ Food [MeSH]
  - > Food and Beverages
- ☐ Climate [MeSH]
  - > Environment ; Atmosphere
- ☐ Lizards [MeSH]
  - > Reptiles

**Add suggested term**

**Suggested Terms ?**

- ☐ Animalia
  - ☐ Chordata
  - ☐ Vertebrata

- Library of Congress Web Archives Minerva project
- Smithsonian Field Notebook project
- US Geological Survey, USGS Thesaurus
- Universidad Carlos III de Madrid (UC3M)
- Inst. Legal Information Theory & Techniques, NRC, Italy

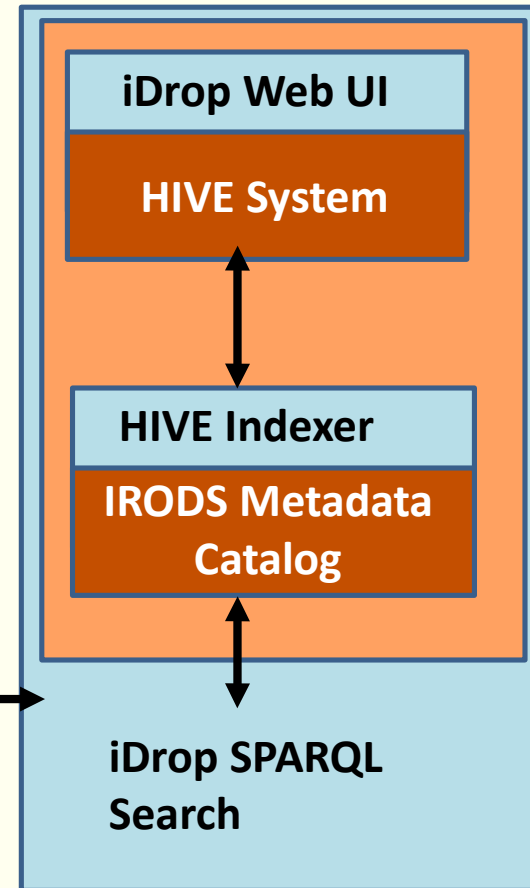
# HIVE/iRODS Integration

## Demo

- <https://vpn.renci.org/dana/home/index.cgi>
- <http://centos6.irods.renci.org:8080/idrop-web2>
  1. Search HIVE
  2. Index with HIVE
  3. Query via HIVE



User uses SPARQL for rich metadata queries, displaying links to DFC files and collections.



# HIVE Across the US DataNets

---

Survey ~ a framework studying controlled vocabulary use across all DataNets

1. Which controlled vocabularies?
2. Purposes that these controlled vocabularies serve (e.g. subject description of datasets or description of analytical processes or protocols that have been applied to certain datasets)
3. Facilitators and inhibitors of controlled vocabulary use by **data contributors, curators, NSF DataNet Partner administrators, and repository infrastructure developers**

[https://unc.qualtrics.com/SE/?SID=SV\\_3fU0xOeRbH6jntb](https://unc.qualtrics.com/SE/?SID=SV_3fU0xOeRbH6jntb).



# Conclusions

---

- Controlled vocabularies encourage consistent classification of data
  - With DFC (Datanet Federation Consortium) we'll be addressing findability of data on distributed grids
- HIVE (or the HIVE approach) allows users to search and apply terms from multiple vocabularies
- Common languages can be generated in different ways
  - Emphasize the benefits, and reduce the limitations
- Acknowledgements: Many people, students, IMLS, NSF, etc.

# Technical overview and architecture

- HIVE combines several open-source technologies to provide a framework for vocabulary services.
- Java-based web services can run in any Java application server.
- Demonstration website @ RENCI and NESCent
- Open-source Google Code (<http://code.google.com/p/hive-mrc/>).

